Journal of Economic Methodology 8:3, 385–414 2001

Discovered preferences and the experimental evidence of violations of expected utility theory

Robin P. Cubitt, Chris Starmer and Robert Sugden

Abstract The discovered preference hypothesis appears to insulate expected utility theory (EU) from disconfirming experimental evidence. It asserts that individuals have coherent underlying preferences, which experiments may not reveal unless subjects have adequate opportunities and incentives to discover which actions best satisfy their preferences. We identify the confounding effects to be expected in experiments, were that hypothesis true, and consider how they might be controlled for. We argue for a design in which each subject faces just one distinct choice task for real. We review the results of some tests of EU which have used this design. These tests reveal the same violations of the independence axiom as other studies have found. We conclude that the discovered preference hypothesis does not justify scepticism about the reality of these effects.

Keywords: discovered preference, common consequence effect, common ratio effect, single-task design

1 INTRODUCTION

That individuals act on stable and coherent preferences is a fundamental assumption in most economic theory. The received model of coherent preferences is provided by expected utility theory (EU), as represented by such axiom systems as that of Leonard Savage (1954). Ever since Maurice Allais' (1953) critique of the 'American school', there has been controversy about the descriptive validity of these axioms. Until the late 1970s, however, criticism was marginalized. Since then, two inter-connected developments have changed the terms of the debate. First, there has been an explosive growth in the use of laboratory experiments in economics. As a result, there is now a large body of experimental evidence which appears to show that individuals' choices among lotteries deviate systematically from the predictions of EU; among the many such regularities that have been found are the violations of the independence axiom predicted by Allais. Second, various alternative theories have been proposed that can accommodate these deviations. So it is no longer possible to defend the use of EU simply by pointing to the absence of any alternative theory.[1] Nevertheless, many economists are still reluctant to conclude that EU has been disconfirmed as a theory of economic decision-

Journal of Economic Methodology ISSN 1350-178X print/ISSN 1469-9427 online
http://www.tandf.co.uk/journals
DOI: 10.1080/1350178011010374 8

making. In this paper, we appraise what, from a methodological viewpoint, seems to be the soundest defence of EU against the experimental evidence: the discovered preference hypothesis.

In the light of the developments we have just described, it seems that any satisfactory defence of EU in economics will have to define the domain of economic theory – or at least, of those branches of the theory that use EU – in such a way as to exclude the experiments in which EU fails. Crude defences of this kind are sometimes put forward, for example the assertion that the purpose of economic theory is to explain behaviour in 'real world' institutions and that evidence from the laboratory is therefore not relevant in assessing the success of the theory. A variant of this argument would allow that some laboratory experiments, in which financially-motivated subjects interact within simplified market-like institutions, count as 'real', while regarding the experiments in which violations of EU have typically been found as outside that domain. These defences seem to depend on a form of instrumentalism which licenses the view that a theory succeeds, just to the extent that its predictions are confirmed *within the domain in which it is intended to be applied*.

We take it that, in the crude and unsupported form in which we have stated it above, this line of argument has no merit. The domain in which a theory can properly be tested cannot defensibly be *identified* with the domain in which it is to be applied. It would be absurd for a civil engineer to deny the relevance of laboratory tests of the strength of bridge girders on the grounds that civil engineering is concerned only with the properties of 'real' bridges. It would be just as absurd for the engineer to accept the results of laboratory tests only if what has been tested is the strength of a complete but simplified bridge. In just the same way, the components of an economic theory may be capable of being tested outside the domain in which the complete theory is intended to be used. In judging whether a theory intended for use in domain D_1 can properly be tested in domain D_2, we have to ask whether there is any *reason* why the theory should work in D_1 but not in D_2. Thus, a putative defence of EU against the apparently disconfirming laboratory evidence must not only define a domain of economic application which excludes the laboratory tests in which the theory fails; it must also offer an explanation of *why* the theory applies to that domain but not to the relevant laboratory environments.

Just such a defence can be constructed from a set of closely related arguments which have been put forward by a number of prominent experimental economists, including Vernon Smith (1989), Glenn Harrison (1994), Charles Plott (1996), and Ken Binmore (1999). From these arguments we shall distil a core hypothesis which, following Plott, we call the *discovered preference hypothesis*.

According to this hypothesis, each individual has coherent preferences, but these preferences are not necessarily revealed in decisions. Facing a particular decision task – whether inside or outside the laboratory – the individual may

not know which of the actions open to him would best satisfy his preferences. This is something that has to be *discovered*, and discovery may involve processes of information gathering, deliberation, and trial-and-error learning. It is only when these processes are complete that the individual's behaviour reveals his true (or *underlying*) preferences. However, it is a crucial feature of the hypothesis that underlying preferences are independent of the particular specification of the discovery process: any process which provides sufficient opportunities and incentives for deliberation and learning will ultimately elicit the same underlying preferences. In this respect, the discovered preference hypothesis is distinct from the hypothesis, proposed by some psychologists, that preferences are *constructed* in the process of decision-making, and that different processes may induce the construction of different preferences (e.g. Payne *et al.* 1992).[2]

In the discovered preference literature, theories of rational choice, such as EU, are interpreted as hypotheses about underlying preferences, rather than about transient choices. Thus, it is argued, experimental tests of such theories are not valid unless subjects have been given adequate opportunities and incentives for discovering how to satisfy their preferences. If this condition does not hold in the experimental designs which have found (supposed) violations of EU, then EU, as interpreted in this literature, has not been disconfirmed.[3]

Notice that this argument does not exclude *all* laboratory experiments from the domain of economic theory. This is perhaps not surprising, given that the argument has been put forward by experimentalists. Nevertheless, this feature of the discovered preference literature is a significant virtue: we would surely be entitled to be suspicious of an argument which purported to show of some theory that no conceivable laboratory test could fall within its domain. So, we might say, the discovered preference hypothesis implies that the boundary of the domain of economic theory runs *through the laboratory* – or, more precisely, through the set of possible experimental designs. An appraisal of the hypothesis, and of its implications for the interpretation of experimental findings, must therefore address the specifics of experimental design.

In the next four sections of this paper, we discuss the implications of the discovered preference hypothesis for the design of experimental tests of EU. We argue that, in this context, the discovered preference hypothesis should be interpreted as an argument for certain kinds of experimental control. We seek to identify the experimental controls that could legitimately be called for by a commentator who accepts the discovered preference hypothesis, and who wishes to assess the descriptive validity of EU in relation to the domain in which that theory is generally applied by economists. We argue that a specific experimental design, the *single-task individual-choice* design, is particularly effective in securing the relevant controls. In an experiment using this design, each subject faces only one distinct choice task; in tests of EU, the task is typically a simple choice between a pair of lotteries.

We shall not be claiming that the single-task individual-choice design is the *only* reliable design for tests of EU. We ourselves have often used other designs, and we stand by our interpretations of the results of those experiments. Nevertheless, we recognize that the discovered preference argument is an important claim, which has been put forward by leading experimental economists. We want to respond to that argument in ways that its proponents can accept, while retaining an interpretation of EU which leaves that theory open to experimental test. In this paper, therefore, we ask whether systematic violations of EU occur in the presence of experimental controls for those confounding effects that would be implied by the truth of the discovered preference hypothesis. In the final section of the paper, we review data from nine experiments which have used the single-task individual-choice design to test the independence axiom of EU. This data set, which records choices made by 977 subjects, gives us a sound basis for assessing the validity of the axiom.

2 SOME PRELIMINARY REMARKS

Many experimental economists will be surprised to read that our arguments will support the single-task individual-choice design, rather than the designs based on repeated markets that are favoured by some of the proponents of the discovered preference hypothesis.

One cause for surprise may be that we are recommending *single-task* designs, when much of the discovered preference literature emphasizes the importance of repetition. But, as we shall make clear later, our concept of 'single task' includes designs which allow a subject repeated practice of the relevant task before facing it for real. It also includes designs in which *the same* task is faced repeatedly. The essential feature of the design is that each subject faces only one *distinct* task (and it is faced for real at least once). Depending on the context, it may or may not be appropriate to include opportunities for practice, or to have repetition.

It may also cause surprise that we are recommending *individual-choice* rather than market-based designs. Advocacy of market-based designs is often supported by references to an apparent contrast between the findings of two of the principal research programmes of experimental economics. Our own position is conditioned by the belief that these programmes investigate fundamentally different questions. It may help to avoid confusion if we explain this belief at the outset.

One research programme investigates the properties of individuals' preferences, as revealed in their decision-making behaviour. The experimental tasks that have been used most commonly in these investigations are ones in which each subject, independently of every other subject, chooses among a pair of options, or reports the amount of some good (usually money) that he regards as equally preferable as some given option. Such experiments have

found many apparently robust and systematic violations of conventional assumptions about preferences, including the violations of EU which are the focus of this paper.[4]

Another programme investigates the properties of market institutions. Here, the most common experimental design is to set up simplified markets in which subjects participate by interacting in closely controlled ways; payoff schemes are designed so as to *induce* particular preferences. (That is, a subject who maximizes monetary payoff will act *as if* motivated by the preferences that the experimenter wishes to induce.) Such experiments have often shown that conventional microeconomic predictions about equilibrium prices and volumes of trade are quite successful, provided that the relevant markets are repeated a sufficiently large number of times.[5]

A superficial contrast between these two sets of results seems to suggest that conventional economic theories receive more support in experiments using market designs than in experiments using individual-choice designs. This has led to speculation that market environments tend to induce the kind of rationality that is assumed in economic theory. Thus, it is suggested, the violations of rationality that are found in individual-choice experiments may have little relevance for behaviour in markets; the discovered preference hypothesis might be taken to rationalize this suggestion.

However, it is important to notice that the main results of the two research programmes do not actually contradict one another. The conventional theories that are disconfirmed in the first programme are theories *about preferences*. The sense of 'rationality' that these theories assume is that of the internal consistency of preferences. In contrast, the conventional theories that are confirmed in the second programme are theories about the equilibrium prices and quantities of trade which, in particular kinds of market, result *from given preferences*. The sense of 'rationality' that these theories assume is that of finding and choosing those actions which, in relatively complex interactive environments, maximally achieve one's objectives – *given that such objectives exist*. There would be no inconsistency in accepting both sets of results at face value, and in concluding that the evidence suggests that people are rational in the second sense but not in the first.

Our purpose here is not to argue for any particular broad-brush interpretation of the findings of experimental economics. We are simply pointing to the difference between the questions asked in these two research programmes. Because of these differences, one should not presuppose that the experimental designs used in one programme are suitable for use in the other. Rather, one should evaluate the discovered preference hypothesis on its own merits, and consider its implications for the investigation of particular questions. That is what we shall do in this paper.

3 DISCOVERED PREFERENCES AND SUBJECT ERROR

The discovered preference hypothesis implies that, for a given individual and a given experimental task, there is a 'correct' response – the response that is consistent with underlying preferences. Failure to give this response is a form of error. Thus, to invoke the discovered preference hypothesis to challenge the validity of a particular experimental design is to claim that, in that design, subjects are liable to make errors.

On our reading of the discovered preference literature, four possible kinds of subject error have been discussed as having the potential to create confounding effects in experiments, whether concerned with tests of EU or more generally. These are:

1 Misunderstanding of experimental procedures: This form of error occurs if a subject fails to understand the task she has been set. The subject may realize that she does not understand, but be unwilling to admit the fact. Or she may *think* she understands what she is expected to do, but her interpretation of the task may be different from that intended by the experimenter. For example, in an induced-preference experimental market, a subject in the role of seller may think that she is required to sell at the highest available price, even if this involves making a loss.

2 Invalid logical inference: A subject may understand how an experimental procedure works, but make incorrect inferences about how best to satisfy her preferences within that procedure. For example, consider an experiment which uses the Becker–De Groot–Marschak (BDM) mechanism[6] to elicit selling prices for some good. The subject might understand how the BDM mechanism works, and be sure how much the good is worth to her, yet fail to understand that it is a weakly dominant strategy to report her true valuation. Instead, she might overstate her valuation, perhaps thinking that this is a shrewd tactical move, or acting on a heuristic that has served her well in other selling situations.

3 Disequilibrium beliefs: In experiments in which subjects interact with one another, what it is best for one subject to do may depend on what other subjects do. In such cases, each subject has to act on subjective beliefs about the actions of the others. Such beliefs are in *equilibrium* if, when all subjects act on their beliefs, their combined actions confirm those beliefs. Disequilibrium beliefs can be interpreted as a form of error. For example, consider an ultimatum game in which the first player to move proposes some division of £10, and the second player either accepts this (in which case, each player receives her proposed share) or rejects it (in which case, neither player receives anything). Suppose the first player believes that the second player would reject any offer less than £2.50, while in fact the second player would accept any positive offer. In this case, the first player's actions reflect his subjective beliefs about the second player's actions, but they do not reflect *true* beliefs.

4 False expectations about affect: When choosing between alternative courses of action, a subject may know exactly what objective consequences would follow from each action, but fail to foresee some of the affective qualities of the subjective experiences which correspond with those consequences. We shall call the relationship between objectively described consequences and the corresponding affective experiences the *consequence-affect relationship*. Failure to foresee this relationship correctly might reflect lack of knowledge of the relevant affective qualities; or it might be the product of some psychological bias which focuses the subject's attention on some aspects of a decision problem at the expense of others.[7] For example, consider a choice between some amount of money with certainty, and a lottery which gives a chance of a larger prize. The subject may underestimate the pain of regret she would feel if she chose the lottery and failed to win.

The fact that an experimental design allows subjects to commit one or more of these kinds of error is not necessarily a valid criticism of that design. All of these types of error have counterparts outside the laboratory and are potentially relevant for explanations of economic behaviour. Thus, an experimenter might want to investigate behaviour in settings in which such errors are possible. Given that we are concerned with the implications of the discovered preference hypothesis for experimental methodology, we suggest that the real issue is that of *experimental control*. That is: for any given investigation, *if the experimental results are to be interpreted on the assumption that a certain type of error does not occur*, then the ideal design would minimize the possibility of that type of error.

In order to draw conclusions from an economic experiment, it is necessary to make background assumptions about how subjects understand the tasks they face. To the extent that the experimenter is uncertain or wrong about subjects' understandings of a task, there is a failure of experimental control. In practice – especially given the principle that subjects should not be deliberately deceived[8] – this kind of control generally requires that subjects' understandings coincide with what the experimenter takes to be the truth. Thus, subjects are normally assumed to understand the relevant experimental procedures, as set out in truthful instructions. And, where this is necessary for their performance of experimental tasks, subjects are assumed to draw valid inferences from those instructions. If for some such experimental design these assumptions do not hold, that design is flawed.

To say this is not to deny the usefulness of experiments which are specifically designed to allow *controlled* tests of subjects' ability to make logical inferences.[9] But in such experiments, the domain in which the validity of subjects' inferences is to be investigated is precisely delimited; outside that domain, subjects are still assumed to understand instructions and to reason correctly about them.

A similar argument applies to disequilibrium beliefs. How individuals form expectations about other people's behaviour is an important issue for experimental research. Controlled experiments can be set up to investigate how, in repeated interactions, subjects' beliefs evolve, and whether these beliefs converge to equilibrium.[10] But, in an experiment whose objective is to elicit individuals' preferences, it is a failure of control if the implications of a subject's responses depend on that subject's beliefs about other subjects' behaviour, and if those beliefs are inaccessible to the experimenter. Thus, when interactive market mechanisms are used to elicit preferences, it is normal to run the relevant markets repeatedly until subjects' behaviour stabilizes, and to assume that, under these conditions of stability, subjects' beliefs are in equilibrium. If for some such experimental design this assumption does not hold, control has failed.

We have argued that, if the interpretation of results depends on the assumption that a particular error does not occur, the ideal design would minimize the possibility of that type of error. Does this argument imply that experiments should be designed so as minimize errors arising from subjects' false expectations about affect? Clearly, the answer depends on the hypothesis that is being tested. More specifically, it depends on whether that hypothesis is or is not to be interpreted on the assumption that individuals, when choosing among consumption bundles or lotteries, hold correct beliefs about the consequence-affect relationship. If the hypothesis to be tested *is* to be interpreted on that assumption, then an experimental procedure which elicited preferences by investigating choices over commodity bundles or lotteries would lack control if the implications of a subject's responses depended on the subject's beliefs about the consequence-affect relationship, and if those beliefs were inaccessible to the experimenter. Analogously with the case of disequilibrium beliefs about other people's actions, one might try to finesse this problem by assuming that subjects hold true beliefs about the consequence-affect relationship. And then, if in some particular experiment that assumption did not in fact hold, there would be a failure of control.

In their discussions of discovered preference, Plott (1996) and Binmore (1999) both seem to propose a tendency for individuals' preferences over commodity bundles and lotteries to become more consistent with EU as those individuals gain experience of consuming the relevant commodities or lottery outcomes. Neither author offers more than a sketch of the supporting argument. Recall, however, that the discovered preference hypothesis requires that, for any given individual, there is a set of 'correct' underlying preferences over actions (i.e. over the commodity bundles or lotteries which might be chosen). We suggest that the most natural way to reconcile Plott's and Binmore's hypothesis about the effects of consumption experience with the logic of discovered preference is to suppose that underlying preferences over actions are determined by the *true* consequence-affect relationship, and that an individual's learning of this relationship is facilitated by repeated

consumption. On this reading, Plott and Binmore are proposing an interpretation of EU in which the theory applies only to underlying preferences that are determined by the true consequence-affect relationship.

If EU is interpreted in this way, valid tests of EU are possible only if there is experimental control over the relationship between *ex post* affect and subjects' *ex ante* beliefs about affect. But the Plott–Binmore interpretation of EU is not the interpretation that is normally used in economics. As both writers acknowledge, their interpretation of 'underlying preference' drastically restricts the domain of conventional economic theory, excluding all those situations in which decision-makers do not correctly predict the affective qualities of consequences. This rules out many applications that most economists would see as entirely uncontroversial. For example, Plott (1996: 226) excludes all 'new tasks', such as buying a house, from the domain of the theory of rational choice, while accepting that such cases 'abound in economics' (see also Binmore 1999: F17). Such an interpretation also threatens to undermine what many would see as part of the *raison d'être* of EU – its status as a theory of choice *under uncertainty*. On the Plott–Binmore account, and if knowledge of the consequence-affect relationship is induced only by repeated consumption experience, EU seems to apply only to decision problems that are repeated a significant number of times; and what is repeated must include not only the act of decision, but also the resolution of any uncertainty and the experience of the resulting outcome. This restriction effectively excludes cases of uncertainty (understood as cases in which 'objective' probabilities or relative frequencies cannot be defined) from the domain of EU. In other words, the whole project of *subjective* EU is called into question.

On what we take to be the conventional interpretation of EU, an individual's preferences are defined over lotteries with objectively-described consequences, such as consumption bundles or amounts of money. These preferences are postulated to satisfy the EU axioms, *irrespective* of the individual's beliefs about the consequence-affect relationship.[11] On the conventional interpretation, the theory can encompass cases in which, as an individual gains experience of consuming a commodity, her preferences *change*; but the EU axioms are assumed to apply to preferences both before and after any such change. Thus, if the object of an experiment is to test EU, as that theory is standardly interpreted, it is not necessary that subjects hold *true* beliefs about the consequence-affect relationship. All that is necessary is that subjects' beliefs about this relationship – whether true or false – do not vary across the experiment in ways which might confound the relevant test. For example, if EU is being tested by a between-subjects comparison of responses to two different choice tasks, there should be no systematic differences in beliefs between the two sets of subjects.

The idea that consistency is not a general property of preferences, but something that is induced by repeated consumption experience, is a novel and potentially important hypothesis, which is capable of being tested

experimentally. But such testing is quite distinct from testing the standard interpretation of EU. This paper is concerned with the latter enterprise. Accordingly, from now on we shall concentrate on the first three of the kinds of subject error considered above – misunderstanding of experimental procedures, invalid logical inferences, and disequilibrium beliefs about other subjects' actions.

4 MINIMIZING SUBJECT ERROR

We focus on three alternative experimental designs. Our concern is with the effectiveness of these designs in reducing subject error in tests of EU.

In the most commonly used design, each subject faces a number of different tasks, each of which requires a choice to be made between a pair of lotteries. We shall call this the *multiple-task individual-choice* design. We use the term 'individual choice' to signify that subjects do not interact with each other; each subject responds, independently of other subjects, to one or more choice tasks. 'Multiple task' signifies that each subject faces two or more *distinct* tasks, each of which may or may not be repeated. Usually, this design is used in conjunction with the *random lottery incentive mechanism*: at the end of the experiment, one task is selected at random, and the subject then plays out the lottery chosen in that task and receives the outcome of that play.

It is often suggested that designs in which subjects participate in markets have greater external validity than do designs in which subjects make simple choices among options (e.g. Smith 1989, 1994, Plott 1996), and there has been a growing tendency for experimentalists to use market designs to test hypotheses about individuals' preferences (e.g. Shogren *et al.* 1994, Cox and Grether 1996, Evans 1997, Myagkov and Plott 1997). In these *market-based* designs, subjects participate in (usually repeated) markets, and hypotheses are tested either by using individual-level data on subjects' market behaviour (e.g. their bids or asking prices in an auction mechanism), or by using group-level data on equilibrium prices.

In this paper, we compare market-based and multiple-task individual-choice designs with the *single-task individual-choice* design. In this latter design, each subject faces only one distinct task, and faces it for real at least once; the task requires the subject to choose one option from a set. Tests of the EU axioms using this design are carried out between subjects. The design has several variants. The simplest is the *one-shot* form: each subject performs the relevant choice task only once, knowing that this choice is for real. In the *with practice* variant, each subject practices taking the relevant decision and playing out the chosen lottery, without reward, at least once (and possibly several times) before making a final decision for real.[12] The experiments we describe in section 4 all use the one-shot form of the design, but, in our methodological discussions, we shall consider both of these variants. For completeness, we add that the single-task individual-choice design also has

with repetition variants, although we know of no experiments that have used them. In these variants, each subject faces the relevant choice task two or more times, with each occasion potentially being for real; the task may be for real on every occasion, or a random lottery mechanism may be used to select one occasion to be for real. In this paper, we shall not discuss these versions of the design.

How effective is each of these designs in reducing subject error? Among the proponents of the discovered preference hypothesis, there seems to be general agreement that the following criteria are particularly important in assessing the tendency of an experimental design to reduce subject error:

1 Transparency/simplicity: Experimental tasks should be as simple and transparent as possible, so as to limit the amount of deliberation and learning that is required of subjects.
2 Incentives: To the extent that subjects' performance of tasks is enhanced by the voluntary expenditure of mental effort on deliberation and learning, subjects should perceive that such effort is adequately rewarded.
3 Learning opportunities: Subjects should have opportunities for trial-and-error learning about the experimental task and about the consequences of alternative strategies that they might adopt.

In the following subsections, we use these criteria to assess the relative merits of different designs. When it is useful to refer to concrete cases, we will use the experiments that are to be reported in this paper to illustrate how the single-task individual-choice design can be used to elicit preferences over lotteries.

4.1 Transparency/simplicity

We shall say that an experimental task is *transparent* to the extent that the nature of the task, and what is expected of the subject in it, is easily understood by the subject. We shall say that a task is *simple* to the extent that, for a subject who understands what is expected of her, performing the task requires little effort. Simplicity and transparency do not necessarily go together. (For example, for a subject who has learned the elements of arithmetic, the task of multiplying 3986 by 9407 without using a calculator is transparent but not simple.)

It is hard to conceive how any task that could be used for eliciting preferences could be more transparent than the single-task individual-choice design, in the one-shot form with pairwise choices. In an experiment using this design, two alternative options are described to the subject, who is then asked to choose one of them. Whichever option she chooses, she is given. That is it.

In the experiments reported in this paper, each option is a lottery defined in terms of a random draw from a set of numbered discs or tickets. There are at most two non-zero money payoffs, each of which is assigned to a block of disc

or ticket numbers. Payoffs are always in whole numbers of UK pounds, and the associated probabilities are simple fractions such as 1/4 or 4/5. The structure of these tasks is similar to that of many familiar decisions in relation to gambling, such as whether or not to buy lottery tickets. Given that the objective is to elicit preferences *over lotteries*, a choice between two such options seems to be just about as simple a task as it is possible to construct. Of course, this is not to claim that subjects never misunderstand such choice problems or find them difficult. In our experience as experimenters, there is no such thing as a task which no potential subject will have any difficulty in understanding or performing. Our concern here is with the *comparative* simplicity and transparency of tasks in different designs.

Multiple-task designs are unambiguously more complex than single-task ones. In multiple-task designs, subjects are presented with several distinct tasks, each of which is comparable in complexity to the one task in a single-task design. If a random-lottery incentive system is used, subjects additionally have to understand the mechanism which selects the task that is for real. The fact that the random-lottery set-up has few salient counterparts outside the laboratory is a further obstacle to its being readily understood.

Designs based on experimental markets are more complicated than individual-choice designs. Participants have to be *taught* the procedures that are to be followed in laboratory markets. For example, participants in a Vickrey auction[13] have to be shown how to submit bids, and how the market price is determined. After a subject has understood these formal procedures, he may still fail to grasp that it is in his interest to bid according to the true value to him of the traded good. Typically, this is learned only after a process of trial and error. If the objective of an experiment is to elicit individuals' preferences, market-based designs interpose an unnecessary layer of complexity between preferences and actions. These complexities are *additional to*, not substitutes for, any difficulties subjects might have in understanding the objects over which their preferences are elicited. Thus, when market-based designs are used to test hypotheses about preferences over lotteries, subjects submit bids to buy or sell those lotteries; understanding how the market works does not eliminate the need to understand the lotteries themselves.

The individual-choice task is not only simple, it also mimics the structure of many everyday economic decision problems. For almost anyone who is likely to be a subject in an economics experiment, the economic decisions of which he will have most experience are those that are made in the role of price-taking consumer: a range of options are on offer, and the individual chooses which of them to take. Most people have much less experience of participating in auction-like markets as active traders. Thus, granted that our object is to elicit the preferences that guide subjects' behaviour in the real economy, it is not at all clear that designs which use auctions or other forms of laboratory market have greater external validity than individual-choice designs.

These remarks are not intended to suggest that either multiple-task or

market-based designs are prohibitively complex, or that they are inappropriate for tests of EU. Nevertheless, it seems indisputable that the single-task individual-choice design is simpler and more transparent than either of the other two.

4.2 Incentives

One of the ways in which subject error can be reduced is through the expenditure of mental effort by the subjects themselves. Subjects can give more or less attention to their instructions; they can deliberate more or less about the options open to them; they can be more or less careful in making logical inferences; and, if they are given opportunities for trial-and-error learning, they can give more or less attention to those opportunities. One of the main functions of incentives is to increase the benefits of mental effort relative to the costs. (Simplicity is the other side of the coin: the simpler the task, the less costly it is to solve the problem of working out what to do.)

For experiments that test EU, incentives have an additional function. According to a number of theories of choice under risk which are rivals to EU, decision-making behaviour is influenced by affective experiences such as anticipations of fear, hope, regret and disappointment. If experimental tests of EU are to be fair to these alternative hypotheses, the nature of the task should be such that these (supposed) affective experiences could arise. The use of significant real payoffs is a way of satisfying this requirement.

When combined with expected payoffs that are sufficient to induce subjects to volunteer to participate in experiments, single-task designs typically offer very favourable benefit/effort tradeoffs. For example, as pointed out in section 4.1, little mental effort is needed to understand the choice tasks in the experiments reported in this paper. The potential rewards to this effort are substantial. There are two kinds of choice task in these experiments: 'scaled down', in which the probabilities of winning the prizes in the relevant lotteries are relatively small, and 'scaled up', in which these probabilities are relatively large. The prizes range from £3.00 to £24.00; the expected values of the lotteries range from £0.90 to £4.80 in scaled-down tasks, and from £3.00 to £19.20 in scaled-up tasks. For the (mostly student) subjects, a relevant comparison is a typical unskilled wage rate of around £4 per hour.

Clearly, for given lotteries, random-lottery designs generate less favourable benefit/effort tradeoffs than do single-task designs, as Harrison (1994) points out. The random-lottery format may also weaken the affective experiences associated with decision-making under risk: the fact that each decision task has only a small chance of being for real may make it harder for the subject to imagine what it would feel like to take on the various lotteries.[14]

In an experiment in which preferences over lotteries are elicited through markets, a subject not only has to decide how much each lottery is worth to him, but also has to choose a trading strategy. Discovering an optimal bidding

strategy for a laboratory auction, whether by deductive reasoning or by trial and error, is not an easy task. Other things being equal, then, the benefit/effort tradeoff will be less favourable in a market experiment than in an individual choice experiment.

4.3 Learning opportunities

In view of the importance attached to repetition by the proponents of the discovered preference hypothesis, it might seem that the single-task design could be recommended only in conjunction with practice or repetition. But rather than accepting this conclusion as it stands, it is necessary to consider *why* repetition has been thought to be so important.

Smith (1989), Plott (1996) and Binmore (1999) all emphasize the role of repetition in *reaching equilibrium* in interactive experiments. In repeated interactive experiments, subjects can use inductive methods to form expectations about one another's behaviour. But this aspect of repetition has no significance for individual-choice designs, since these involve no interaction between subjects.

Smith, Plott and Binmore suggest two further arguments in favour of repetition, which apply to non-interactive as well as to interactive experiments. The first is that repetition is a mechanism for instructing subjects about experimental procedures and for eliminating misunderstandings. This consideration undoubtedly counts in favour of repetition or practice, and applies very generally. But it has less force the more transparent the experimental procedures are. Given clear instructions, it seems unlikely that many subjects would misunderstand a single-task experiment which required no more than a choice between two single-stage lotteries. It may be significant that the experimentalists who put most stress on repetition tend to be those who specialize in market experiments (as Smith and Plott do) or in game-theoretic experiments (as Binmore does). In these more complex experimental environments, the problem of ensuring that subjects receive adequate instruction assumes greater importance.

The second argument is that repetition allows subjects to try out different responses, and thus to use trial-and-error learning as a supplement to logical inference in discovering how best to satisfy their preferences within an experimental design. For example, consider again the case of a subject in an experiment that uses the BDM mechanism to elicit selling prices for some good. Suppose the subject does not understand that it is a dominant strategy to report her true valuation, but instead overstates her valuations. After experience of the BDM mechanism, she might learn to follow the dominant strategy – with or without understanding the reasoning which recommends that strategy. But could a similar argument support repetition or practice *for choice tasks*? Given clear instructions, one would not expect a subject to have to engage in trial-and-error learning in order to realize that if she faces a single

choice between two lotteries A and B, and if she prefers A, then her best strategy is to choose A. Just as in the case of instruction, then, the value of repetition or practice as a means of facilitating learning depends on the complexity of the experimental environment.

It might also be argued that the repetition of tasks could allow subjects to revise their expectations about the relationship between consequence and affect. However, we have argued that in the context of tests of EU, as standardly interpreted, such revisions should be understood as changes in preferences, and not as the correction of subject error. There may well be differences between individuals' responses to one-shot and repeated decision problems. Both kinds of situation are important in the real world that economists seek to explain, and are legitimate subjects for experimentation.[15] We take it that, on a standard interpretation, EU has implications for both kinds of decision. Experiments which (like those to be reported in section 6) test EU by using the single-task design in its one-shot form must be understood as testing hypotheses about individuals' behaviour in one-off decision problems. By using the single-task design *with practice*, it might also be possible to pick up some of the consequences of revisions of expectations about affects.

5 CONTAMINATION

The single-task design has an additional advantage over multiple-task and market-based designs: the avoidance of cross-subject and cross-task contamination. In this section, we explain why this is so. We focus on tests of the independence axiom of EU, but our argument has much more general application.

Cross-subject contamination occurs when an experimental design treats the responses of different subjects as independent observations, when in fact the responses of one subject are influenced by those of another. Cross-task contamination occurs when, in a within-subject test, a subject's responses to different tasks are treated as independent observations of his preferences, when in fact his response to one task is influenced by his response to another. Both kinds of contamination can generate spurious confirmations (or indeed violations) of theories such as EU, which impose conditions of internal consistency on preferences.

The independence axiom of EU imposes a consistency property on preferences: for certain pairs of lotteries $\{R, S\}$, $\{R', S'\}$, each individual's preference ranking of R in relation to S must be the same as his ranking of R' in relation to S'. Tests of the axiom typically work by comparing revealed preferences over two such pairs. For such a test to be valid, observations of choices over $\{R, S\}$ and over $\{R', S'\}$ must be independent of one another. If instead there is contamination between the two sets of observations, apparent consistency between those observations could be the product of that contamination.

This danger is particularly acute in multiple-task experiments that involve repetition or practice. Repetition and practice give opportunities for trial-and-error learning, which can be desirable if subjects are liable to make errors in the early stages of an experiment. But, if the discovery of underlying preferences is not to be confounded with effects that are due to contamination, any learning processes that precede supposedly independent observations must not interact with one another.

In EU, as in all conventional theories of rational choice, preferences over given options are *context-independent*. That is, each individual has a single preference ordering over the set of all conceivable options, and this ordering governs her choices in every decision problem. Context independence is important in the application of EU: for example, it allows us to make predictions about behaviour in one context from observations made in another. It is also an essential assumption in all comparative-static analysis. If, for a given subject and a given task, the learning process converges to different revealed preferences depending on how other subjects behave or on what other tasks the experiment contains, that process is context-dependent. In multiple-task experiments, such context-dependent learning processes may generate spurious confirmations of EU.

Market-based designs which use repetition are vulnerable to a particular version of this problem. As such an experiment progresses, subjects learn about the terms on which other subjects are willing to trade (for example, by observing the transactions made by other traders in open-cry double auctions,[16] or by observing market prices in successive Vickrey auctions). Thus, at the same time as subjects' preferences are being elicited, they are being given information about the decisions made by their fellow-subjects. This information may influence subjects' subsequent behaviour. For example, if individuals have some tendency to imitate one another, participants in an auction may be induced to revise their bidding behaviour so as to make it more similar to the behaviour of others (or to what, from the information at their disposal, they infer that behaviour to be). Of course, the provision of price information would be entirely appropriate in an experiment whose object was to investigate the efficiency of market institutions, given agents' induced preferences. In that case, the availability of the information is simply a property of the institution under investigation. But, if the object of the experiment is to elicit actual preferences and to test them for consistency, price information is a potential source of contamination.

That the provision of price information could introduce cross-subject contamination has not usually been acknowledged. For example, James Cox and David Grether (1996) investigate preference reversal, eliciting valuations by using a second-price auction mechanism. Their finding that the rate of preference reversal falls as the auction is repeated has been quoted by both Smith (1994: 118) and Plott (1996: 231) as evidence that repeated market experience tends to induce behaviour that conforms with standard

assumptions about preferences. But, Cox and Grether also find that subjects' bids are positively correlated with previous market prices. They interpret this effect as the result of between-subject transfers of information, and as one of the principal factors explaining their results (p. 400).[17] This effect might be the result of mutual imitation; it is certainly evidence of interaction between the learning processes of different subjects.

Another form of spurious confirmation of EU can occur if, within subjects, the learning processes associated with different decision tasks interact with one another. In principle, any multiple-task design which tests EU within subjects is vulnerable to this kind of contamination. As such an experiment progresses, a subject may learn to use particular modes of analysis or heuristics. We might expect the heuristics that a subject uses in tackling the later tasks in the experiment to be those that are in some way adapted to, or suggested by, the earlier tasks. If (as is standard practice in multiple-task experiments) the order of tasks is randomized independently for each subject, the heuristics used for any given task are liable to be influenced by the nature of each of the other tasks.

In practice, this form of contamination seems most likely to be a problem when there are many tasks in an experiment and when there are close similarities between those tasks. Under such circumstances, subjects have the opportunity (and, to the extent that mental effort is costly, the incentive) to learn to use simplifying heuristics. These could be highly specific to a particular set of tasks. If the different tasks in an experiment are tackled by means of common heuristics, we should expect to find properties of cross-task consistency in subjects' responses. But, this is not necessarily evidence that subjects are discovering *context-independent* preferences that have those consistency properties. Whether that is true depends on whether the sets of heuristics that are applied to given tasks are context-independent, or whether they are conditioned by the combination of tasks in the experiment. If the object of the experiment is to test hypotheses about the consistency of individuals' preferences, the possibility of this kind of cross-task contamination must be a concern.

As an example which raises just these problems of interpretation, consider the following multiple-task design which has been used, in conjunction with the random lottery incentive system, in several recent experimental tests of EU.[18] The tasks are choices between pairs of lotteries. There is a large number of such tasks, and each is faced more than once. However, all the lotteries in these tasks are similar to one another in that they are different probability mixes of the same three or four money prizes. The data collected are used to estimate stochastic models of preferences. The results of these experiments are remarkably similar: as each experiment progresses, subjects' responses converge towards a pattern that can be represented by a stochastic form of EU. Utility functions in this EU model are closely approximated by constant relative risk aversion functions of the form $u = ax^{\alpha}$, where u is utility, x is

monetary gain in the task, and a and α are positive constants; the value of α is of the order of 0.3, which indicates a very high degree of risk aversion. At first sight, these results may seem to support the hypothesis that individuals have EU preferences, which are discovered only after sufficient decision-making experience. Notice, however, that the property of constant relative risk aversion with respect to monetary gains implies that subjects' decisions are entirely explained by the relative sizes and relative probabilities of the payoffs in each task. This striking regularity in behaviour is difficult to reconcile with conventional EU, in which there is a context-independent utility function whose domain is defined over *levels* (as contrasted with *increments*) of wealth.[19] We cannot eliminate the possibility that the regularity is induced by context-dependent heuristics which are learned in the course of these experiments.

When multiple-task experiments use the random-lottery incentive system, further forms of cross-task contamination can occur. For example, Charles Holt (1986) suggests that subjects may treat such an experiment as a single decision task, reducing complex lotteries to simple ones by the calculus of probability. If subjects behave in this way, the common consequence effect[20] will not be observed in random-lottery experiments, irrespective of whether or not preferences satisfy the independence axiom (Starmer and Sugden 1991): thus, there can be spurious confirmations of EU. In fact, the few tests that have been carried out have found no significant evidence of cross-task contamination for pairwise choices in random-lottery experiments (Cubitt *et al.* 1998a); but the possibility of such contamination makes this design less than ideal.

We conclude that, other things being equal, it is desirable that experimental tests of EU use designs which screen out both cross-subject and cross-task contamination. The single-task individual-choice design clearly satisfies this condition, while market-based and multiple-task designs do not.

6 SINGLE-TASK TESTS FOR THE COMMON CONSEQUENCE AND COMMON RATIO EFFECTS

In the preceding sections, we have considered alternative experimental designs for tests of the EU axioms. We have argued that the single-task individual-choice design is superior to its main rivals on grounds of transparency, simplicity, incentives, and the avoidance of cross-subject and cross-task contamination. We now turn to the evidence that has been generated by a programme of experiments that have used variants of this design.

Over the last decade, we and our associates Jane Beattie and Graham Loomes have conducted a series of experiments to investigate the validity of various experimental designs for testing hypotheses about individuals' preferences. Our experiments have typically required some operational definition of 'true' preferences, against which the validity of particular

preference-elicitation methods can be assessed. Taking the preferences elicited by the single-task individual-choice design as canonical in this sense, we have used adaptations of that design as controls in our experiments. As a result, we have accumulated a considerable body of data on individuals' choices in single-task designs.

In this paper, we analyse data from a set of nine experiments. This set consists of every experiment conducted up to the time of writing (June 2000) by us, or in research projects in which we have participated, which uses a single-task design, or a near variant, and which tests for *common consequence* or *common ratio* effects in choices over single-stage lotteries.

The common consequence and common ratio effects were first postulated by Maurice Allais (1953). Let a, b be money consequences such that $a > b > 0$. Consider the lotteries $R_1 = (a, \lambda; 0, 1 - \lambda)$, $R_2 = (a, \lambda p; b, 1 - p; 0, [1 - \lambda]p)$, $R_3 = (a, \lambda p; 0, 1 - \lambda p)$, $S_1 = S_2 = (b, 1)$, $S_3 = (b, p; 0, 1 - p)$, such that $0 < \lambda < 1$ and $0 < p \leq 1$. The independence axiom implies that the preference ranking of R_1 with respect to S_1 is the same as that of R_2 with respect to S_2, and as that of R_3 with respect to S_3. In each of the choice tasks $\{R_i, S_i\}$, R_i is the *riskier* option and S_i is the *safer* option. The common consequence effect is a tendency for preferences over the pair $\{R_3, S_3\}$ to be less risk averse than preferences over $\{R_2, S_2\}$; the common ratio effect is a tendency for preferences over $\{R_3, S_3\}$ to be less risk averse than preferences over $\{R_1, S_1\}$.

The single-task design for testing for the common consequence effect works as follows. Subjects are divided at random into two groups. Each subject in one group faces the *scaled-up* single-choice task $\{R_2, S_2\}$; each subject in the other faces the *scaled-down* task $\{R_3, S_3\}$. The null hypothesis is that subjects' preferences satisfy the independence axiom. Since the two groups are random samples from a common population, the null hypothesis implies that the probability that any subject in the scaled-up group chooses R_2 is equal to the probability that any subject in the scaled-down group chooses R_3. The alternative hypothesis, i.e. that there is a common consequence effect, implies that the second probability is greater than the first. The test for the common ratio effect follows the same principles, but using $\{R_1, S_1\}$ as the scaled-up task in conjunction with the scaled-down task $\{R_3, S_3\}$.

These tests can be interpreted as assuming that each individual has a strict and non-stochastic preference between each pair of lotteries. But they can also be interpreted in a different way, which allows for stochastic variation by means of a *random preference* model. In such a model, there is for each individual a probability distribution over possible preference relations, each of which satisfies the restrictions of the relevant theory (in the case of our null hypotheses, the theory is EU). A subject's response to an experimental task is taken to be governed by a particular preference relation, drawn at random from this distribution.[21]

As explained above, each of the nine experiments we report was carried out as part of some larger investigation. Here, we report only the results of the

tasks relevant to the independence axiom. Each subject faced one of these tasks and once only. Insofar as these experiments tested the independence axiom, they did so as a by-product of testing other hypotheses. In consequence, there are many differences of detail between the designs of the nine experiments. In fact, although all the experiments use the individual-choice format, only one of them uses the single-task design in the pure form described in section 4. The designs used in the other experiments diverge from that ideal type in various ways. These divergences are described below. We suggest that, in most cases, they are quite minor. However, the reader is welcome to form his or her own judgment, separately for each experiment, about the extent to which these divergences compromise the merits of the single-task design.

The experiments are listed in the first column of table 1. As the second column shows, experiments CC1 to CC3 test for common consequence effects, while CR1 to CR6 test for common ratio effects. CC3, CR5 and CR6 are the most recent experiments, and are reported for the first time. Full details of the other experiments can be found in the relevant publications.

The remaining columns of the table summarize the most salient design features of the experiments. The third column describes the subject pool for each experiment. 'UEA' and 'Sussex' denote that subjects were recruited during teaching terms on the campuses of, respectively, the University of East Anglia and the University of Sussex. The subjects recruited were mostly students; in age, gender and subject of study, they were broadly representative of the student populations of those universities. 'UEA summer school' denotes that subjects were recruited on the UEA campus during the summer vacation; in this case, subjects were mostly mature students attending Open University summer schools. No one took part in more than one experiment.

The fourth column indicates the display used to describe the two lotteries to the subjects, and the way in which consequences in the two lotteries were juxtaposed in the (implicit or explicit) act/event matrix. Three different displays were used. In all cases, lotteries were described by assigning money prizes to states of the world, corresponding to numbered tickets or discs, one of which was to be drawn at random. In the *strip* display, the two lotteries are shown separately. Figure 1a illustrates this display for an experiment in which each subject played out his chosen lottery by drawing a disc from a bag of 100 discs. The *matrix* display represents a choice between lotteries as an act/event matrix, the rows representing lotteries, the columns representing sets of states of the world, and the entries in the cells describing the corresponding consequences. Figure 1b illustrates this display. In this case, each subject was given oral instructions which told him that he had to choose one of the options '*c*' or '*d*', that the numbers along the top of the matrix referred to numbered lottery tickets, that the entries in the matrix were amounts of money (in UK pounds), and that the numbers at the bottom of each column showed the relevant number of chances in 100. The *text* display describes the two lotteries separately, in words. An example of this display is shown in Figure 1c.

Table 1 Design details

Experiment	Effect investigated	Subject pool	Display/ juxtaposition	Location of task
CC1 (Starmer and Sugden, 1991: Groups A, D)	common consequence	UEA summer school ($n = 80$)	matrix/disjoint	follows 21 hypothetical choice tasks (including matched common consequence task)
CC2 (Cubitt, Starmer and Sugden, 1998a: Groups 1.3, 1.4)	common consequence	UEA ($n = 82$)	strip/overlap	follows 19 unrelated hypothetical choice tasks
CC3 (new experiment)	common consequence	UEA ($n = 160$)	strip/overlap	follows unrelated random-lottery experiment
CR1 (Beattie and Loomes, 1997: Groups 3, 5)	common ratio	Sussex ($n = 96$)	text/overlap	follows unrelated experiment with hypothetical tasks
CR2 (Cubitt, Starmer and Sugden, 1998a: Groups 2.1, 2.2)	common ratio	UEA ($n = 97$)	strip/overlap	follows 19 unrelated hypothetical choice tasks
CR3 (Cubitt, Starmer and Sugden, 1998a: Groups 3.1, 3.2)	common ratio	UEA ($n = 105$)	strip/overlap	combined with 19 hypothetical choice tasks in random order (including matched common ratio task)
CR4 (Cubitt, Starmer and Sugden, 1998b: Groups 1, 5)	common ratio	UEA ($n = 102$)	text/overlap	no other tasks
CR5 (new experiment)	common ratio	UEA ($n = 160$)	strip/overlap	follows unrelated random-lottery experiment
CR6 (new experiment)	common ratio	UEA ($n = 95$)	strip/overlap	follows unrelated random-lottery experiment

Lottery 1:		Lottery 2:	
discs 1–25	discs 26–100	discs 1–20	discs 21–100
you win £5	you win nothing	you win £8	you win nothing

1a: Strip display (as used in scaled-down task of CR5)

	1 – 25	26 – 80	81 – 100
c	7.00	0.00	0.00
d	0.00	0.00	10.00
	25	55	20

1b: Matrix display (as used in scaled-down task of CC1)

Choose either Option A or Option B:

Option A The controller will draw a chip from the bag. If it is numbered 1–80, you will receive nothing. If it is numbered 81–100, you will receive £16.

Option B The controller will draw a chip from the bag. If it is numbered 1–75 you will receive nothing. If it is numbered 76–100, you will receive £10.

1c: Text display (as used in scaled-down task of CR4)

Figure 1 Sample displays

In the scaled-up tasks, S_1 and S_2 are degenerate lotteries, and so the juxtaposition of consequences in these tasks is uniquely determined. But in the scaled-down tasks, the consequences of the two lotteries can be juxtaposed in various ways. This juxtaposition can be described by the probability that the consequence of S_3 is b, conditional on the consequence of R_3 being a. In the *overlap* design, this probability is 1; in the *disjoint* design, it is zero. Regret theory permits common ratio and common consequence effects for the disjoint case, but not for the overlap case (Starmer and Sugden 1989). Thus, the overlap design controls for one factor – the influence of regret – which might contribute to common consequence and common ratio effects in the disjoint design.

The final column of table 1 describes how, for each subject, the single-choice task in each experiment is located in relation to other parts of that experiment. The variety of entries in this column reflects the variety of purposes for which the experiments were run. CR4 is the experiment with the pure single-task design: in this experiment, each subject faced only one task, knowing that this task was for real.

In CC2, CR1 and CR2, each subject first responded to a series of *hypothetical* tasks (choices among lotteries in the case of CC2 and CR2, valuation tasks in CR1). These hypothetical tasks were the same (except for statistically independent randomization) for all subjects in the relevant experiment, and were unrelated to the common consequence or common ratio task which, in the final part of the experiment, was faced for real. While not quite as elegant as the pure single-task design, these designs share its main merits: transparency and simplicity of the relevant choice task, strong incentives, and control of cross-subject and cross-task contamination. These designs differ from the pure design only in respect of contamination: whilst they certainly eliminate cross-subject contamination, they eliminate cross-task contamination only if it is assumed that unrelated hypothetical tasks are non-contaminating.

CC1 and CR3 were similar to the experiments described in the previous paragraph, in that each subject faced a number of hypothetical choice tasks and one real task. Again, the hypothetical tasks were the same for all subjects in a given experiment, except for independent randomization. In these experiments, however, each subject faced both the scaled-up and the scaled-down forms of the relevant task. If the scaled-up task was faced for real, the scaled-down task was faced as one of the hypothetical tasks; and vice versa. These designs do not fully control for cross-task contamination. However, we conjecture that the special salience of a task that is for real, combined with the large number and variety of hypothetical tasks in these experiments, reduces the likelihood that responses to the real task were influenced by the presence of the matched hypothetical task.

CC3 and CR5 are two components of a single experiment, which had two distinct parts. In the first part, each subject faced a set of tasks (pairwise choices over combinations of money and chocolates) which were unrelated to the task to be faced in the second part. These first-part tasks were presented in a random-lottery design. Having completed these tasks, but before discovering which of them was for real, each subject faced one common consequence or common ratio task *for real*; subjects took away from the experiment the sum of their payoffs from its two parts. In the present context, it is a limitation of CC3 and CR5 that the set of first-part tasks was not the same for all subjects. The 320 subjects in the combined experiment were divided into eight subgroups, each of which faced a different set of first-part tasks. Then, in the second part of the experiment, two subgroups faced the scaled-up common consequence task, two faced the scaled-down common

consequence task, and so on. Between each pair of subgroups which faced the same second-part task, there was no significant difference in second-part responses. Nevertheless, we cannot rule out the possibility of contamination between the two parts of this experiment.[22]

CR6 is a component of an experiment designed as a follow-up to that described in the previous paragraph. It was identical with CR5 in all but two respects. The preceding tasks (choices between money and chocolates) were slightly different; and in this case the set of first-part tasks *was* the same, except for independent randomization, for both scaled-up and scaled-down groups.[23]

The parameters used in the various experiments are shown in Table 2 (payoffs are in UK pounds). Although the absolute sizes of the payoffs differ considerably, the values of a/b, λ and p are similar in all eight experiments. The results of the three common consequence experiments are shown in Table 3; the results of the six common ratio experiments are shown in Table 4. For each experiment, the number and percentage of subjects choosing each option in the scaled-up and scaled-down tasks are shown. The final column shows the z-statistic for a test of the hypothesis that the proportion of subjects choosing the riskier lottery is higher in the scaled-down task than in the scaled-up one. (Negative values indicate that the difference in proportions is in the direction implied by the alternative hypothesis, i.e. that a common consequence or common ratio effect is present. A single asterisk denotes that, in a one-tail test, this difference is significant at the 5 per cent level; double asterisks denote significance at the 1 per cent level.) In eight out of the nine experiments, the proportion of risky choices is greater in the scaled-down task; in four of these cases the difference is significant at the 1 per cent level and in two other cases it is significant at the 5 per cent level.

The final row in each table pools the data from all the relevant experiments and reports the corresponding z-statistic. In each case, the proportion of subjects who choose the riskier lottery is greater in the scaled-down tasks, and this difference is significant at the 1 per cent level. These pooled tests should be treated with care. Their formal validity depends on the assumption that the

Table 2 Parameters

Experiment	a	b	a/b	λ	p
CC1	10.00	7.00	1.43	0.80	0.25
CC2	10.00	6.00	1.67	0.75	0.33
CC3	4.00	3.00	1.33	0.83	0.30
CR1	15.00	10.00	1.50	0.80	0.25
CR2	24.00	15.00	1.60	0.80	0.25
CR3	15.00	10.00	1.50	0.80	0.25
CR4	16.00	10.00	1.60	0.80	0.25
CR5	8.00	5.00	1.60	0.80	0.25
CR6	8.00	5.00	1.60	0.80	0.25

Table 3 Results of common consequence experiments

Experiment	Number (%) choosing R_2	Number (%) choosing S_2	Number (%) choosing R_3	Number (%) choosing S_3	z
CC1	13 (32.5)	27 (67.5)	23 (57.5)	17 (42.5)	−2.25**
CC2	25 (65.8)	13 (34.2)	24 (54.5)	20 (45.5)	1.04
CC3	28 (35.0)	52 (65.0)	49 (61.2)	31 (38.8)	−3.32**
total	66 (41.8)	92 (58.2)	96 (58.5)	68 (41.5)	−3.01**

Table 4 Results of common ratio experiments

Experiment	Number (%) choosing R_1	Number (%) choosing S_1	Number (%) choosing R_3	Number (%) choosing S_3	z
CR1	7 (14.6)	41 (85.4)	26 (54.2)	22 (45.8)	−4.08**
CR2	21 (41.2)	30 (58.8)	24 (52.2)	22 (47.8)	−1.08
CR3	14 (28.6)	35 (60.7)	25 (44.6)	31 (55.4)	−1.70*
CR4	19 (38.0)	31 (62.0)	25 (48.1)	27 (51.9)	−1.03
CR5	25 (31.3)	55 (68.8)	47 (58.8)	33 (41.3)	−3.50**
CR6	23 (50.0)	23 (50.0)	33 (67.3)	16 (32.7)	−1.72*
total	109 (33.6)	215 (66.4)	180 (54.4)	151 (45.6)	−5.34**

participants in the experiments were drawn independently from the same subject pool. Rather than make the unsupported claim that this strong assumption is true, we prefer to treat the pooled z-tests merely as summary statistics: they give some indication of the weight of evidence in support of the hypothesis that the common consequence and common ratio effects occur.

When these results are viewed together, there is very strong evidence that decision-making behaviour deviates systematically from the implications of the independence axiom, in the direction that is consistent with the common consequence and common ratio effects. We recognize that all but one of these experiments use designs which diverge in various ways from the ideal type of the single-task design. Nevertheless, if our arguments in support of that design are valid, these data are among the most reliable experimental data in existence that can be used to test EU as a theory of one-off choices. The most parsimonious explanation of these results is surely that individuals' preferences violate the independence axiom of EU in a systematic way.

7 CONCLUSION

The object of this paper was to appraise an argument which has been used to defend expected utility (EU) theory, as used in economics, against the criticism that experimental evidence shows the predictions of the theory to be systematically disconfirmed. The core of this argument is the discovered preference hypothesis.

We have argued that, in the context of experimental economics, the discovered preference hypothesis should be interpreted as an argument for certain kinds of experimental control. We have identified the confounding effects that might be expected to occur, if that hypothesis were true, and we have examined the power of different experimental designs to control for these effects. We have argued that the strongest controls for these effects are provided, not (as some experimental economists have suggested) by market-based designs, but by the single-task individual-choice design. Economists who take the discovered preference argument as a reason for scepticism about the experimental evidence against EU ought therefore to give particular weight to the results of single-task individual-choice experiments.

We have appraised the empirical validity of the independence axiom of EU by using data from nine experiments which have used this design, or close variants of it. Taken as a whole, these data provide strong evidence of the existence of the common consequence and common ratio effects in non-repeated choices. We conclude that the discovered preference hypothesis does not justify scepticism about the reality of those effects. Although, it is sometimes suggested that only repeated choices constitute valid tests of EU, we have argued that this view is mistaken: EU has implications for behaviour in both repeated and non-repeated tasks. Whether violations of EU become less frequent as subjects gain experience remains an open question.

ACKNOWLEDGEMENTS

The research reported in this paper was carried out as part of the Economic and Social Research Council's programme on Risk and Human Behaviour (award L 211 252 053). Robert Sugden's work was also supported by the Leverhulme Trust. We thank Roger Backhouse, Charles Plott and participants in seminars at Caltech and the University of Oxford for comments, and Jan Anderson, Ian Bateman and Alistair Munro for help in organizing the experiments.

Robin P. Cubitt, University of East Anglia
r.cubitt@uea.ac.uk
Chris Starmer, University of Nottingham
chris.starmer@nottingham.ac.uk
Robert Sugden, University of Each Anglia
r.sugden@uea.ac.uk

NOTES

1 These developments in experimental research and in theory are reviewed by Starmer (2000).
2 Plott (1996: 227–8) makes exactly this distinction between discovered and constructed preferences.
3 Neither Smith, Plott, nor Binmore explicitly claim that the classic experiments which have found violations of EU are invalid as tests of underlying preferences. However, Smith and Plott argue for the superiority of market-based experimental designs over individual-choice designs, while Binmore (1999: F17) rejects experiments which do not involve repetition. These arguments implicitly challenge the validity of the classic tests of EU, which use individual-choice designs without repetition.
4 The findings of this programme are surveyed by Camerer (1995).
5 The claim that the findings of this programme are broadly supportive of conventional microeconomic theory is made by Smith (1989) and Plott (1991).
6 In an experiment using the BDM mechanism, each subject reports the minimum price at which he is willing to sell a good which he owns. The price that will actually be paid is determined by a random device. Each subject then sells at this randomly determined price if, and only if, it is no less that his stated minimum.
7 Plott (1996: 226–7) seems to have this second possibility in mind – and, in particular, a bias which focuses attention on the immediate at the expense of the distant – when he says: 'Untutored choices reflect a type of myopia. The individual is purposeful and optimizing, but exhibits limited awareness about the immediate environment or the possible longer-run consequences of any acts that might be taken. Responses are "instantaneous" or "impulsive" . . .'
8 The prevailing view among most experimental economists is that subjects in economics experiments should not be deceived. For a discussion of this issue, see the contributions on this topic in the *Journal of Economic Psychology* (1998).
9 The 'selection task', due to Wason (1968), is an example of an experimental paradigm for investigating subjects' capacities for logical reasoning.
10 The experiment reported by Van Huyck *et al.* (1990) is an example of this kind of investigation.

11 Arguably, EU as conventionally applied to lotteries might be interpreted as the reduced form of a fuller theory of choice under uncertainty, in which consequences are defined in terms of affects. On this interpretation, the validity of EU as a reduced form depends, not on the assumption that individuals' beliefs about the relationship between (objective) consequences and affects are *true*, but on the assumption that the EU axioms apply to preferences in the fuller theory.
12 Two of the current authors have used the with-practice design in a recent investigation of sequential decision-making (Cubitt and Sugden 2001).
13 In a Vickrey auction, the potential buyers of an object make simultaneous sealed bids; the highest bidder takes the object, paying an amount equal to the second highest bid.
14 Loewenstein and Adler (1995) find evidence which supports a similar hypothesis in relation to the psychology of loss aversion.
15 See Camerer (1996), Starmer (1999) and Loewenstein (1999) for arguments against the claim that economics should be concerned only with repeated decisions.
16 In an open-cry double auction, buyers and sellers simultaneously call out offers to buy or sell at stated prices; at any time, anyone can accept anyone else's offer.
17 Knetsch *et al.* (1999) argue, as we do, that the provision of price information in repeated auctions is cross-subject contamination. They report experimental results which suggest that bids in Vickrey auctions are influenced both by observations of past prices and by expectations of future prices.
18 The experiments we are referring to are described by Hey and Orme (1994), Loomes and Sugden (1998) and Hey (1999). Our analysis of them here draws on panel data analysis of these three experiments presented by Loomes *et al.* (2001) and Moffatt (2000).
19 A similar problem is revealed by Rabin (2000) who presents a 'calibration theorem' showing that EU cannot reconcile stylized facts about behaviour over 'small' and 'large' wealth intervals.
20 This classic violation of the independence axiom is defined in section 4 below.
21 This model of stochastic variation is explained by Loomes and Sugden (1998).
22 The reader may ask why subgroups were not constructed independently for the two parts of the experiment. The answer is that the second-part tasks were incidental to the main purpose of the experiment; they were added as means of supplementing the rewards given to subjects, so as to aid recruitment. For two of the subgroups, first-part payoffs could be expected to be particularly low, and so these subgroups were assigned to the 'best' second-part task, the scaled-up common ratio task. Thus, the results of CR5 may be contaminated by a wealth effect. As the other six subgroups had broadly similar first-part payoffs, this problem is much less likely to be significant for CC3. Further details of this experiment are available from the authors.
23 In the follow-up experiment, no subject had a low expected first-part payoff, so the problem of providing adequate rewards, described in footnote 22, did not arise. Further details of this experiment are available from the authors.

REFERENCES

Allais, M. (1953) 'Le comportment de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école americaine', *Econometrica,* 21: 503–46.
Beattie, J. and Loomes, G. (1997) 'The impact of incentives upon risky choice experiments', *Journal of Risk and Uncertainty* 14: 149–62.
Binmore, K. (1999) 'Why experiment in economics?', *Economic Journal*, 109: F16–24.

Camerer, C.F. (1995) 'Individual decision making'. In J.H. Kagel and A.E. Roth (eds) *Handbook of Experimental Economics,* Princeton, NJ: Princeton University Press, pp. 587–703.

Camerer, C.F. (1996) 'Rules for experimenting in psychology and economics, and why they differ'. In W. Albers, W. Güth and E. Van Damme (eds) *Experimental Studies of Strategic Interaction: Essays in Honor of Reinhard Selten*, Berlin: Springer-Verlag.

Cox, J.C. and Grether, G.M. (1996) 'The preference reversal phenomenon: response mode, markets and incentives', *Economic Theory* 7: 381–405.

Cubitt, R.P., Starmer, C. and Sugden, R. (1998a) 'On the validity of the random lottery incentive system', *Experimental Economics* 1: 115–31.

Cubitt, R.P., Starmer, C. and Sugden, R. (1998b) 'Dynamic choice and the common ratio effect: an experimental investigation', *Economic Journal* 108: 1362–80.

Cubitt, R.P. and Sugden, R. (2001) 'Dynamic decision-making under uncertainty: an experimental investigation of choice between accumulator gambles', *Journal of Risk and Uncertainty* 22: 103–28.

Evans, D.A. (1997) 'The role of markets in reducing expected utility violations', *Journal of Political Economy* 105: 622–36.

Harrison, G.W. (1994) 'Expected utility and the experimentalists', *Empirical Economics* 19: 223–53.

Hey, J.D. (1999) 'Does repetition improve consistency?', *Mimeo*, University of York.

Hey, J.D. and Orme, C. (1994) 'Investigating generalizations of expected utility theory using experimental data', *Econometrica* 62: 1291-326.

Holt, C.A. (1986) 'Preference reversals and the independence axiom', *American Economic Review* 76: 508–15.

Knetsch, J.L., Tang, F.-F. and Thaler, R.H. (1999) 'The endowment effect and repeated market trials: is the Vickrey auction demand revealing?', Working Paper, University of Chicago.

Loewenstein, G. (1999) 'Experimental economics from the viewpoint of behavioural economics', *Economic Journal* 109: F25–34.

Loewenstein, G. and Adler, D. (1995) 'A bias in the prediction of tastes', *Economic Journal* 105: 929–37.

Loomes, G., Moffatt, P. and Sugden, R. (2001) 'A microeconometric test of alternative stochastic theories of risky choice', *Journal of Risk and Uncertainty*, forthcoming.

Loomes, G. and Sugden, R. (1998) 'Testing different stochastic specifications of risky choice', *Economica* 65: 581–98.

Moffatt, P. (2000) 'Microeconometric modelling of the role of experience in decision-making under risk', *Mimeo*, University of East Anglia.

Myagkov, M. and Plott, C.R. (1997) 'Exchange economies and loss exposure: experiments exploring prospect theory and competitive equilibria in market environments', *American Economic Review* 87: 801–28.

Payne, J.W., Bettman, J.R. and Johnson, E.J. (1992) 'Behavioral decision research: a constructive processing perspective', *Annual Review of Psychology* 42: 87–131.

Plott, C.R. (1991) 'Will economics become an experimental science?', *Southern Economic Journal* 57: 901–19.

Plott, C.R. (1996) 'Rational individual behaviour in markets and social choice processes: the discovered preference hypothesis', In K.J. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds) *The Rational Foundations of Economic Behaviour*, Basingstoke: Macmillan, pp. 225–50.

Rabin, M. (2000) 'Risk aversion and expected-utility theory: a calibration theorem', *Econometrica* 68: 1281–292.

Savage, L. (1954) *The Foundations of Statistics*, New York: Wiley.

Shogren, J.F., Shin, S.Y., Hayes, D.J. and Kliebenstein, J.B. (1994) 'Resolving

differences in willingness to pay and willingness to accept', *American Economic Review* 84: 255–70.
Smith, V.L. (1989) 'Theory, experiment and economics', *Journal of Economic Perspectives* 3: 151–69.
Smith, V.L. (1994) 'Economics in the laboratory', *Journal of Economic Perspectives* 8: 113–31.
Starmer, C. (1999) 'Experimental economics: hard science or wasteful tinkering?', *Economic Journal* 109: F5–15.
Starmer, C. (2000) 'Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk', *Journal of Economic Literature* 38: 332–82.
Starmer, C. and Sugden, R. (1989) 'Probability and juxtaposition effects: an experimental investigation of the common ratio effect', *Journal of Risk and Uncertainty* 2: 159–78.
Starmer, C. and Sugden, R. (1991) 'Does the random-lottery incentive system elicit true preferences? An experimental investigation', *American Economic Review* 81: 971–8.
Van Huyck, J.B., Battalio, R.C. and O. Beil, R. (1990) 'Tacit coordination games, strategic uncertainty, and coordination failure', *American Economic Review* 80: 234–48.
Wason, P.C. (1968) 'Reasoning about a rule', *Quarterly Journal of Experimental Psychology* 20: 273–81.